Assessing the Approximate Validity of Moment Restrictions*

Pascal Lavergne†

May 1, 2020

Abstract

We propose a new theoretical framework to assess the *approximate* validity of overidentifying moment restrictions. Their validity is evaluated by the divergence between the true probability measure and the closest measure that imposes the moment restrictions of interest. The divergence can be chosen as any of the Cressie-Read family. The considered *alternative* hypothesis states that the divergence is smaller than some user-chosen tolerance. Tests are constructed based on the minimum empirical divergence that attain the local semiparametric power envelope of invariant tests. We show how the tolerance can be chosen by reformulating the hypothesis under test as a set of admissible misspecifications. Two empirical applications illustrate the practical usefulness of the new tests for providing evidence on the potential extent of misspecification.

Keywords: Hypothesis testing, Semiparametric models.

JEL Codes: C12, C14, C52.

*I acknowledge financial support from SSHRC and from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program). I thank seminar participants, and in particular Vadim Marmer and Laurent Davezies, for useful comments.

† *Toulouse School of Economics. Email: lavergnetse@gmail.com* Address correspondence: Toulouse School of Economics, 1 Esplanade de l'Université, 31080 Toulouse Cedex 06, FRANCE.

1 Introduction

The goal of this work is to develop "classical" tests for assessing the *approximate* validity of overidentifying restrictions. Since the seminal work by White (1982), it has been widely recognized that misspecification is the rule rather than the exception, and a growing literature has aimed at accounting for potential misspecification in inference. Several authors have adopted a local to zero approach for studying misspecifications of moment restrictions, e.g. instruments that locally violate exogeneity, see Berkowitz, Caner, and Fang (2008, 2012), Bugni, Canay, and Guggenberger (2012), Conley, Hansen, and Rossi (2012), Kraay (2012), Guggenberger (2012), Nevo and Rosen (2012), and Caner (2014). Recent work focuses on consequences for inference on parameters. Andrews, Gentzkow, and Shapiro (2017) and Andrews, Gentzkow, and Shapiro (2018) study sensitivity of estimators to local misspecifications, while Armstrong and Kolesár (2018) and Bonhomme and Weidner (2018) propose robust confidence intervals.

Practically, our approach recognizes that any model is misspecified to some extent, and aims at confirming that misspecification is relatively small. To develop such tests, a central issue is how to measure the extent of misspecification. Here we build on recent work on Generalized Empirical Likelihood (GEL), which include Empirical Likelihood (EL), see Imbens (1993) and Qin and Lawless (1994), Exponential Tilting (ET), see Imbens (1993) and Kitamura and Stutzer (1997), and the Continuously Updated Estimator (CUE-GMM), see Hansen, Heaton, and Yaron (1996) and Antoine, Bonnal, and Renault (2007). As explained by Kitamura (2007), these estimators rely on minimizing a divergence (or contrast) between the distribution of the observations and one that imposes the moment restrictions. We choose as a measure the theoretical Cressie-Read divergence, which has a natural information-theoretic interpretation. This choice is mainly motivated by invariance considerations. Indeed, any measure of validity (or lack of) should not vary if moment restrictions are reformulated in a different but equivalent way. Such a measure should also be invariant to any (potentially nonlinear) reparameterization. As our main instance, we focus on the chi-square divergence, which, for moment restrictions of the form $\mathbb{E}\, g(X,\theta_0) = 0$, measures the extent of misspecification as

$$\min_{\Theta} \mathbb{E}\ \left(g'(X,\theta)\right) \left[\text{Var}\ g(X,\theta)\right]^{-1} \mathbb{E}\ \left(g(X,\theta)\right)\ . \tag{1.1}$$

Clearly, this is measure is zero unless there is overidentification. As will be shown, any other Cressie-Read divergence yields approximately the same theoretical measure of validity if the restrictions are close to be valid. Given a misspecification measure for our moment conditions, we consider as our *alternative hypothesis* that this divergence is smaller than some user-chosen *tolerance*.

The interest of approximate hypotheses has been long recognized in statistics, see e.g. Hodges and Lehmann (1954). As stated by Cox (1958), "exact truth of a (point) null hypothesis is very unlikely except in a genuine uniformity trial." Leamer (1998) argues that "genuinely interesting hypotheses are neighborhoods, not points. No parameter is exactly equal to zero; many may be so close that we can act as if they were zero." Good (1981) and Berger and Delampady (1987) in statistics or McCloskey (1985) in economics, among many others, also advocate for approximate hypotheses. We chose to consider the approximate validity of the moment condition as the *alternative* hypothesis to reflect where the burden of proof is. This is known in biostatistics as *equivalence testing*, see Lehmann and Romano (2005) and the monograph of Wellek (2003). Applications of approximate hypotheses and equivalence testing can be found for example in Romano (2005) and Lavergne (2014) for restrictions on parameters, and in Rosenblatt (1962) and Dette and Munk (1998) for specification testing. With reference to equivalence testing in biostatistics, our tests are labeled *model equivalence tests for moment restrictions.*

Our test is based on the empirical analog of the divergence (1.1). The alternative hypothesis is accepted for small values of the empirical divergence, and the critical value is not derived under the assumption that the moment restrictions are valid. The new tests have interesting properties, in particular they attain the semiparametric power envelope of invariant tests for our hypotheses. Our framework builds on Lavergne (2014), who has focused on restrictions on parameters in parametric models and a Kullback-Leibler divergence. We significantly extend it to assessing the approximate validity of moment restrictions in semiparametric models using a large class of divergences. Our new tests allow to conclude that the model may be misspecified to an extent that is acceptable by the practitioner, as measured by the chosen tolerance. Our work is related to the recent literature on robust inference under local misspecification cited above. For instance, Armstrong and Kolesár (2018) consider $\mathbb{E}\, g(X, \theta_0) = c/\sqrt{n}$ and devise confidence intervals that are robust to such such local misspecifications for c in a user-chosen set $\mathcal{C}$. By contrast our test aims to confirm that misspecification is indeed local and bounded by the user-chosen tolerance, so our approach is complementary to theirs.

The tolerance can be interpreted as a squared percentage, and its square root as the distance of overidentifying restrictions to zero in standard deviations units. Its role is similar to the one of the threshold used for defining weak instruments in Stock and Yogo (2005), who deemed instruments weak if the bias of the IV estimator in standardized units exceed a certain percentage. In a theoretical demand model, Chetty (2012) also measures the degree of optimization frictions (i.e. the extent of model misspecification) through the average utility cost as a percentage of expenditures. Christensen and Connault (2019) similarly use Cressie-Read divergences to measure misspecification and evaluate sensitivity of counterfactuals from an economic model. It can be useful for the researcher to return to the natural units of the application and to assess using expert judgment what the chosen tolerance implies for a particular model. The restatement of model equivalence in terms of overidentifying restrictions that we derive is instrumental in this respect. For instance, in an IV model, it is possible to state how much endogeneity, that is how much correlation between the error term and the instruments, is allowed by choosing a specific tolerance. Finally, it is also possible to let the tolerance vary so as to determine the minimal allowable misspecification that yields to declare model equivalence. Again this can be reinterpreted in terms of local misspecification of the moment restrictions by the researcher to decide whether the model under scrutiny is only slightly or grossly misspecified, as illustrated in Section 3.

The outcome of a model equivalence test is not defining a confidence region of a special kind. A confidence region is a random set such that we are confident with some predetermined level, say 95%, that the true parameters lie in this set. A model equivalence hypothesis defines instead a set such that the probability of falsely concluding that the parameters are in this set is bounded by a small number, say 5%. So the two sets are constructed by controlling different probabilities. Another possible approach would be to rely on power evaluation of overidentification tests. Andrews (1989) proposes approximations of the asymptotic inverse power function of Wald tests for restrictions on parameters as an aid to interpret non significant outcomes. While such an approach might be generalized to overidentification tests, this has not been investigated up to date.[1] To sum up, model equivalence tests for moment restrictions deliver a new type of inference that is complementary to existing methods.

In Section 2, we develop our testing framework first based on the chi-square divergence then on a general Cressie-Read divergence, that includes as special cases the ones used in EL and ET. We show that all divergences are approximately equal for an "almost" correctly specified model, so that the chosen divergence should not matter as soon as the tolerance is small. In Section 3, we illustrate the usefulness of the new tests on two selected empirical examples. In Section 4, we derive the semiparametric envelope of tests that are invariant to transformations of the moment restrictions and we show that our tests reach this envelope. Section 5 concludes. Section 6 contains the proofs of our results.

2 The Tests

2.1 Testing Framework: Chi-Square Divergence

For a random vector $X \in \mathbb{R}^q$ with probability distribution P, we want to assess some implicit restrictions of the form

$$\exists\, \theta_0 \in \Theta \text{ such that } \ \mathbb{E}\, g(X, \theta_0) = \mathbf{0}\,, \tag{2.2}$$

where $g(\cdot, \theta)$ is a m-vector function indexed by a finite-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, $p < m$. To do so, we can evaluate the divergence between P and a measure that imposes these restrictions. Consider the chi-square divergence (or contrast) between two measures Q and P defined as

$$D_2(Q, P) = \mathbb{E}\, \frac{1}{2}\left(\frac{dQ}{dP} - 1\right)^2 = \frac{1}{2}\int \left(\frac{dQ}{dP} - 1\right)^2 dP\,,$$

where $\frac{dQ}{dP}$ denotes the Radon-Nikodym derivative. Hence $D_2(Q, P) \geq 0$ with equality if and only if $Q = P$ $\;P-$almost surely. Twice the chi-square divergence measures the

[1] Wald tests are not invariant to nonlinear transformations of restrictions under scrutiny, see e.g. Gregory and Veall (1985). Moreover, *evaluating* the asymptotic power of a significance test of given level does not directly provide evidence in favor of the approximate validity of the restrictions under consideration. Other issues surround post-experiment power calculations, as summarized by Hoenig and Heisey (2001).

expected squared proportional difference between distributions and is thus an expected squared percentage. For a particular value of $\theta \in \Theta$, let

$$\mathcal{M}_\theta = \left\{ Q \text{ finite measure} : \, Q << P, \, \int dQ = 1, \, \int g(X, \theta) \, dQ = 0 \right\}$$

and $D_2(\mathcal{M}_\theta, P) = \inf_{Q \in \mathcal{M}_\theta} D_2(Q, P)$. A minimizer Q_θ of $D_2(Q, P)$ over $\mathcal{M}_\theta$, if it exists, is labeled a projection of P on $\mathcal{M}_\theta$. Now let $\mathcal{M} = \cup_{\theta \in \Theta} \mathcal{M}_\theta$. A minimizer $Q_\mathcal{M}$ of $D_2(Q, P)$ over $\mathcal{M}$ is a projection of P on $\mathcal{M}$. The quantity

$$D_2(\mathcal{M}, P) = \inf_{\Theta} \, D_2(\mathcal{M}_\theta, P) \tag{2.3}$$

provides a global measure of the approximate validity of the restrictions (2.2). By definition, this measure is invariant to any reparameterization and any transformation of the restrictions. In particular, for any $q \times q$ matrix $A(\theta)$ which is nonsingular for any θ with probability one, the moment restrictions (2.2) remains unaltered if $g(\cdot, \theta)$ is replaced by $A(\theta) g(\cdot, \theta)$, and so does $D_2(\mathcal{M}, P)$. Moreover, a duality approach, as discussed e.g. by Kitamura (2007) and briefly outlined in the supplementary material, shows that

$$D_2(\mathcal{M}, P) = \frac{1}{2} \min_{\Theta} \mathbb{E} \, \left(g'(X, \theta) \right) \left[\text{Var } g(X, \theta) \right]^{-1} \mathbb{E} \, \left(g(X, \theta) \right) \, , \tag{2.4}$$

see Antoine et al. (2007). This is the theoretical objective function used in the CUE-GMM method. Hence twice the divergence has a pretty intuitive content: it measures the square distance to zero of the moment restrictions in standard deviations units.

To assess the approximate validity of our moment restrictions, we consider the alternative hypothesis that $D_2(\mathcal{M}, P)$ is smaller than some tolerance chosen by the practitioner. That is, there is a measure imposing the moment restrictions which is close enough to the true probability measure. We write our alternative hypothesis as

$$H_{1n} : \; 2 \, D_2(\mathcal{M}, P) < \frac{\delta^2}{n} \, .$$

This hypothesis is labeled the *model equivalence hypothesis*. It allows for some local misspecification of the moment restrictions, as apparent from (2.4). The null hypothesis is the complement of the alternative, that is

$$H_{0n} : \; 2 \, D_2(\mathcal{M}, P) \geq \frac{\delta^2}{n} \, .$$

The vanishing tolerance δ^2/n, which makes the alternative hypothesis shrinks, is a purely theoretical but useful device, acknowledging that misspecification is small in a substantive sense, as considered by Romano (2005), Berkowitz et al. (2012), Bugni et al. (2012), Caner (2014), Lavergne (2014), Andrews et al. (2017), Andrews et al. (2018), Armstrong and Kolesár (2018), and Bonhomme and Weidner (2018), among others. In practice, as in our subsequent illustrations, a small but fixed tolerance Δ^2 is typically chosen, where Δ can be seen as a percentage, so one can set $\delta^2 = n\Delta^2$ to run the test. But because the fixed tolerance is small, the asymptotics under a drifting tolerance will approximate the finite sample behavior of the test statistic better than the asymptotics under a fixed tolerance.

2.2 Testing Procedure

With at hand a random sample $\{X_i, i = 1, \ldots n\}$ from X, the empirical divergence of interest is

$$D_2(Q, P_n) = \mathbb{E}_n \frac{1}{2}\left(\frac{dQ}{dP_n} - 1\right)^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(Q(X_i) - 1\right)^2 ,$$

where $\mathbb{E}_n$ denotes expectation with respect to the empirical distribution P_n. Let

$$\mathcal{M}_{n,\theta} = \left\{Q \text{ finite measure}: \; Q << P_n, \; \int dQ = 1, \; \int g(X,\theta)\, dQ = 0\right\}$$

$\mathcal{M}_n = \cup_{\theta\in\Theta}\mathcal{M}_{n,\theta}$, and

$$D_2(\mathcal{M}_n, P_n) = \inf_{\Theta} \inf_{Q\in\mathcal{M}_{n,\theta}} D_2(Q, P_n). \tag{2.5}$$

This quantity is the empirical equivalent of the theoretical divergence and thus provides a natural estimator of the latter. In addition, duality extends to the empirical chi-square divergence, so that

$$D_2(\mathcal{M}_n, P_n) = \frac{1}{2}\min_{\Theta} \mathbb{E}_n \left(g'(X,\theta)\right)\left[\mathrm{Var}_n \; g(X,\theta)\right]^{-1} \mathbb{E}_n \left(g(X,\theta)\right) ,$$

where Var_n denotes the empirical variance, see e.g. Antoine et al. (2007). As a by-product, we obtain the CUE-GMM estimator of the solution of (2.4), which is the value of θ_0 that fulfills (2.2) when the restrictions hold. By contrast to standard two-step GMM, estimation is one-step and does not require a preliminary estimator. The empirical divergence is also invariant to any reparameterization and any transformation of the restrictions, which may not be the case for the two-step GMM optimal objective function, see e.g. Hall and Inoue (2003).

The empirical divergence provides a natural basis for testing H_{0n} against H_{1n}. When the theoretical divergence $2\,D_2(\mathcal{M},P)$ equals $\frac{\delta^2}{n}$, $2n\,D_2(\mathcal{M}_n,P_n)$ converges in distribution to a $\chi^2_r(\delta^2)$, the non-central chi-square with $r=m-p$ degrees of freedom and noncentrality parameter δ^2. The model equivalence test is then defined as

$$\pi_n = \mathbb{I}\left[2\,n\,D_2(\mathcal{M}_n,P_n) < c_{\alpha,r,\delta^2}\right] \,,$$

where c_{α,r,δ^2} is the α-quantile of a $\chi^2_r(\delta^2)$. The test concludes that moment restrictions are approximately valid if the test statistic $2\,n\,D_2(\mathcal{M}_n,P_n)$ is relatively small. This stands in contrast to an overidentification test, which rejects the exact validity of moment restrictions for large values of the test statistic, and for which the critical value is the $1-\alpha$ quantile of a central chi-square distribution. This is because our model equivalence test does not assume that moment restrictions hold under the null hypothesis, as the test aims at confirming that these restrictions approximately hold. While critical values are non-standard, they can be readily obtained from most statistical softwares.

The main properties of the test are easily derived. First, it is invariant to reparameterization and to transformation of the moment restrictions. Second, when $2\,D_2(\mathcal{M},P)$ is large, which corresponds to grossly misspecified restrictions, the test will fail to reject H_{0n} in favor of model equivalence. This can be deduced from the convergence of $D_2(\mathcal{M}_n,P_n)$ to the theoretical divergence $D_2(\mathcal{M},P)$, see Broniatowski and Keziou (2012, Theorem 5.6). In Section 4, we will establish asymptotic optimality of the test.

The objective function based on the chi-square divergence is similar to the GMM one, both at the theoretical and empirical level. Reformulating the problem in terms of the two-step GMM theoretical objective function would yield to write the null and alternative hypotheses in terms of

$$\frac{1}{2}\min_{\Theta}\mathbb{E}\ \left(g'(X,\theta)\right)\left[\mathrm{Var}\ g(X,\theta_1)\right]^{-1}\mathbb{E}\ \left(g(X,\theta)\right) \,, \tag{2.6}$$

with $\theta_1 = \arg\min_{\Theta}\|\mathbb{E}\ g(X,\theta)\|$. Aside the non-invariance of this theoretical criterion, this seems an awkward way to measure the extent of misspecification because $\mathbb{E}\ g(X,\theta)$ is scaled by the standard deviation of $g(X,\theta_1)$. Of course, this should not matter much if the model is only lightly misspecified, but we cannot assume at the outset what we would like to show. For these reasons, we do not aim to extend our analysis to the two-step GMM context. Routines to implement CUE-GMM are now available for many econometric softwares, such as Stata, or languages, such as Gauss, Matlab, or R. In our applications of Section 4, the two-step GMM criterion was found to be pretty close to the Generalized Empirical Likelihood (GEL) ones and thus would yield similar outcomes if used to run a model equivalence test. This does not preclude however the possibility to obtain contradictory outcomes in some other applications.

2.3 Alternative Formulation

We now show how to formulate and interpret the model equivalence hypothesis in terms of overidentification restrictions. As will be seen, such an alternative formulation is intuitive and appealing from an empirical viewpoint. For any $p \times m$ matrix L with full rank p, consider the partition of $g(\cdot, \theta)$ into a p-vector $g_1(\cdot, \theta) = Lg(\cdot, \theta)$ and the remaining $(m-p)$ vector $g_2(\cdot, \theta) = Mg(\cdot, \theta)$, where $[L, M]$ is full rank. Define

$$D_W(\mathcal{M}, P) = \frac{1}{2} \mathbb{E}\, g_2'(X, \theta^*) \Sigma^{-1} \mathbb{E}\, g_2(X, \theta^*)\,,$$

where Σ is the semiparametric efficiency bound on the $\sqrt{n}$-variance for estimating $\mathbb{E}\, g_2(X, \theta^*)$. We will show that this divergence is locally equivalent to $D_2(\mathcal{M}, P)$ in the following sense.

Definition 2.1 *Two divergence measures d_i, $i = 1, 2$, are locally equivalent under a drifting sequence of probability distributions $\tilde{P}^{(n)}, n \geq 1$, if whenever $d_1(\mathcal{M}, \tilde{P}^{(n)}) = o(1)$ or $d_2(\mathcal{M}, \tilde{P}^{(n)}) = o(1)$, we have $d_1(\mathcal{M}, \tilde{P}^{(n)}) = d_2(\mathcal{M}, \tilde{P}^{(n)})(1 + o(1))$.*

Let us introduce the following assumptions.

Assumption 2.1 *(a) Θ is compact; (b) $\mathrm{Var}\, g(X, \theta)$ is positive definite for any $\theta \in \Theta$; (c) For any $p \times m$ matrix L with full rank p, there exists a unique solution θ^* to the equations $L\mathbb{E}\, g(X, \theta) = \mathbf{0}$; (d) $\tilde{\theta}_0 = \arg\inf_\Theta D_2(\mathcal{M}_\theta, P)$ is unique; and (v) $\nabla_\theta \mathbb{E}\, g(X, \tilde{\theta}_0)$ is full rank.*

Assumption 2.2 *Each component of the function $g(\cdot, \theta)$ is twice continuously differentiable in θ over Θ.*

Lemma 2.1 *Under any drifting sequence of probability distributions $\tilde{P}^{(n)}$ such that Assumptions 2.1 and 2.2 hold, $D_2(\mathcal{M}, \tilde{P}^{(n)})$ and $D_W(\mathcal{M}, \tilde{P}^{(n)})$ are locally equivalent.*

Therefore, when $2\, D_2(\mathcal{M}, \tilde{P}^{(n)}) < \delta^2/n$, $D_W(\mathcal{M}, \tilde{P}^{(n)})$ is bounded by $(\delta^2/n)\,(1 + o(1))$. But a test based on a sample of size n would not be able to distinguish a variation in divergence of an order smaller than $1/n$. This entails that the alternative hypothesis H_{1n}, for all practical purposes, is asymptotically the same as (and indistinguishable from)

$$\mathbb{E}\, g_2'(X, \theta^*)\Sigma^{-1}\mathbb{E}\, g_2(X, \theta^*) < \frac{\delta^2}{n}\,. \tag{2.7}$$

This alternative formulation uses a divergence that focuses on the closeness to zero of $m-p$ overidentifying moments in standard deviations units evaluated at θ^*. Moreover, this is independent of the particular choice of the subset $g_2(\cdot, \cdot)$. If there is one degree of overidentification only, i.e. $m - p = 1$, then the above expression becomes

$$|\mathbb{E}\, g_2(X, \theta^*)| < \frac{\delta\sigma}{\sqrt{n}}\,,$$

where σ^2 is the semiparametric efficiency bound for estimating $\mathbb{E}\, g_2(X, \theta^*)$. With a consistent estimator of σ, or of Σ in the general case, one can then evaluate the content of the model equivalence hypothesis in terms of closeness to zero of the overidentification restrictions, and if their number is small, the set defined by (2.7) can be easily graphed. The last formulation is simple and intuitive, but it must be kept in mind that direct tests of this hypothesis would generally not be invariant. We will therefore use this asymptotically equivalent formulation for interpretative purposes only, see Section 3.

2.4 Cressie-Read Divergence Based Test

We here detail the more general tests based on Cressie-Read divergences and we discuss their relationship with the test described in the previous section. As done by Smith (1997), Imbens, Spady, and Johnson (1998), Newey and Smith (2004), and Kitamura (2007), we focus here on the class of divergences based on the Cressie and Read (1984) family of functions

$$\begin{aligned}
\varphi_\gamma(x) &= [x^\gamma - \gamma x + \gamma - 1] \,/\, [\gamma(\gamma - 1)]\,, \quad \gamma \in \mathbb{R}\backslash\{0, 1\}\,, \\
\varphi_1(x) &= x \log x - x + 1, \\
\varphi_0(x) &= -\log x + x - 1\,.
\end{aligned}$$

If $\varphi_\gamma(\cdot)$ is not defined on $(-\infty, 0)$, as for $\gamma = 0$, or when it is not convex on $(-\infty, 0)$ as $\varphi_3(x)$, we set it to $+\infty$ on $(-\infty, 0)$. Hence, all considered functions are strictly convex, positive, and twice differentiable on their domain. The way we wrote the Cressie-Read family of functions slightly differs from most of the econometric literature, but yields the normalization $\varphi_\gamma(1) = 0$, $\varphi'_\gamma(1) = 0$, and $\varphi''_\gamma(1) = 1$, so that all functions behave similarly around 1 up to second-order. For each γ, the Cressie-Read divergence between two measures Q and P is defined as

$$D_\gamma(Q, P) = \mathbb{E}\ \varphi_\gamma\left(\frac{dQ}{dP}\right) = \int \varphi_\gamma\left(\frac{dQ}{dP}\right) dP\,.$$

The quantity $D_\gamma(\mathcal{M}, P) = \inf_\Theta D_\gamma(\mathcal{M}_\theta, P)$ thus provides an alternative global measure of the validity of the moments restrictions (2.2). The cases $\gamma = 1$ and 0 correspond to Kullback-Leibler-type divergences, $\gamma = 1/2$ yields the Hellinger divergence, see Kitamura, Otsu, and Evdokimov (2013), and $\gamma = 2$ the chi-square divergence considered above. The model equivalence hypothesis based on $D_\gamma(\cdot,\cdot)$ writes

$$H_{1n}:\ 2\, D_\gamma(\mathcal{M}, P) < \frac{\delta^2}{n}\,,$$

and the null hypothesis is

$$H_{0n}:\ 2\, D_\gamma(\mathcal{M}, P) \geq \frac{\delta^2}{n}\,.$$

The corresponding empirical divergence is

$$D_\gamma(\mathcal{M}_n, P_n) = \inf_\Theta \inf_{Q \in \mathcal{M}_{n,\theta}} D_\gamma(Q, P_n)\,. \tag{2.8}$$

For $\gamma = 1$, respectively $\gamma = 0$, one obtains as a by-product the exponential tilting (ET) estimator, respectively the empirical likelihood (EL) estimator. The model equivalence test writes

$$\pi_n = \mathbb{I}\left[2\, n\, D_\gamma(\mathcal{M}_n, P_n) < c_{\alpha,r,\delta^2}\right]\,,$$

with the same critical values as the test based on the chi-square divergence. Irrespective of the choice of the divergence, the test retain the same basic characteristics than the test based on the chi-square divergence. In particular, it remains invariant to any transformation of the moment restrictions. But because of the degree of freedom in the choice of the specific divergence, there is a multiplicity of implied model equivalence hypotheses and tests.

We now show that all Cressie-Read divergences are equivalent for locally misspecified models, so that the choice of the divergence should not matter much in practice. A similar but slightly different result has been independently derived by Andrews et al. (2018).

Assumption 2.3 *(a) For any* $\theta \in \Theta$, $D_\gamma(\mathcal{M}_\theta, P) < \infty$. *(b)* $\tilde{\theta}_0 = \arg\inf_\Theta D_\gamma(\mathcal{M}_\theta, P)$ *exists and is unique.*

Lemma 2.2 *For any* γ, *under any drifting sequence of probability distributions* $\tilde{P}^{(n)}$ *such that Assumptions 2.1, 2.2, and 2.3 hold,* $D_\gamma(\mathcal{M}, \tilde{P}^{(n)})$ *and* $D_2(\mathcal{M}, \tilde{P}^{(n)})$ *are locally equivalent.*

Our result entails that the choice of the particular divergence is asymptotically irrelevant for the definition of the model equivalence hypothesis H_{1n}, while there may be some supplementary (theoretical or practical) reason to favor a specific divergence in a particular application. As a result, the alternative formulations of model equivalence derived for the chi-square divergence in Section 2.3 extend to any Cressie-Read divergence. Hence (2.7) is an asymptotically equivalent formulation of model equivalence, irrespective of the chosen divergence. Also the tolerance can be interpreted as a squared percentage or as the square of the distance to zero of the moment restrictions in standard deviations units.

To show the asymptotic equivalence between different Cressie-Read divergences, we use duality, see Kitamura (2007). The strength of the duality principle is that dual optimization is finite-dimensional and concave. For duality to apply, one needs a projection to exist, which is ensured by Assumption 2.3 (a). Basically, this requires that for each θ a measure $Q \in \mathcal{M}_\theta$ exists such that $\frac{dQ}{dP}(x)$ lies in the interior of the support of $\varphi_\gamma(\cdot)$. The projection of P on $\mathcal{M}_\theta$ is then essentially unique, see Keziou and Broniatowski (2006) for more detailed conditions on the existence and uniqueness of projections. This is explicitly assumed in Assumption 2.3 (b). Our technical assumption may seem pretty innocuous in practice. Indeed, one can always restrict the parameter space to the set of θ for which a finite empirical divergence obtains. However it may not be so when moment restrictions are misspecified. Take any function $\varphi_\gamma(\cdot)$ with domain $(0, \infty)$, such as the ones used for EL or ET. The projection measure Q that solves $D_\gamma(\mathcal{M}, P) = D_\gamma(Q, P)$ should be a probability measure with the same support as P. But, in case of misspecification, such a measure may not exist. Issues of GEL estimation methods under misspecification have been documented in the literature. In particular, Schennach (2007) shows that the EL estimator can have an atypical behavior when moment restrictions are invalid, as a projection does not generally exist when the functions in $g(\cdot,\cdot)$ are unbounded. Sueishi (2013) points out that under misspecification there may exist no probability measure in $\mathcal{M}$ with a finite divergence $D_1(\mathcal{M}, P)$. By contrast, because $\varphi_2(\cdot)$ has domain $\mathbb{R}$, and since $\mathcal{M}_\theta$ includes signed measures, a solution always exists when minimizing the chi-square divergence.

2.5 Choice of the Tolerance

The choice of the tolerance $\Delta^2 = \delta^2/n$ used to define model equivalence is key. From the definition of the divergence, and our alternative formulation of Lemma 2.1, the square root of the tolerance is a percentage or equivalently a number of standard deviations units of the moment restrictions. It is similar to the threshold used, for instance, by Stock and Yogo (2005) to characterize weak instruments, or by Chetty (2012) to evaluate the extent of model misspecification. Any analysis of locally misspecified models is met with the choice of a tolerance. For a parametric model, Bonhomme and Weidner (2018) determine a tolerance by choosing the probability of a model detection error based on likelihood ratios. For moment restrictions models, they suggest using specification testing, as do Armstrong and Kolesár (2018). In their applications, Andrews et al. (2017) consider the effect on estimation of local misspecification of several (unscaled) moments, each taken at a time, with a tolerance of $1/n$. Armstrong and Kolesár (2018) perform a sensitivity analysis letting the amount of misspecification depends on the number of potentially invalid instruments. Our above formulation of model equivalence in terms of overidentifying restrictions allow the researcher to return to the application and to asses using expert judgment what the chosen tolerance implies for a particular model, as we will illustrate below. For instance, in an IV model, it is possible to state how much endogeneity, that is how much correlation between the error term and the instruments, is allowed by choosing a specific tolerance.

Tolerance should ultimately be tailored to the specific application at hand. If one does not wish to choose a tolerance at the outset, we may let it vary for a given level of the test. Formally, let

$$\delta^2_{\text{inf}}(\alpha) = \inf\left\{\delta^2 > 0 : 2\,n\;D_2\left(\mathcal{M}_n, P_n\right) < c_{\alpha,r,\delta^2}\right\}\;. \tag{2.9}$$

Hence $\Delta^2_{\text{inf}}(\alpha) = \delta^2_{\text{inf}}(\alpha)/n$ determines the minimal allowable misspecification that yields the test to declare model equivalence.[2] This provides a useful benchmark against which a practitioner may decide a posteriori whether it is a small enough misspecification. Again it can be reinterpreted in terms of moment restrictions to help the researcher reaching a decision. We will illustrate in our applications how this provides valuable information on the model approximate validity.

3 Empirical Illustrations

We here apply our model equivalence tests to two selected empirical problems. This will help us to discuss the choice of the tolerance and the interpretation of the outcomes. All computations used the R package `gmm`, see Chaussé (2010).

3.1 Social Interactions

Graham (2008) shows how social interactions can be identified through conditional variance restrictions. He applies this strategy to assess the role of peer spillovers in learning using data from the class size reduction experiment Project STAR. His model yields conditional restrictions of the form

$$\mathbb{E}\left[\rho(Z_c, \tau^2(W_{1c}), \gamma_0^2)|W_{1c}, W_{2c}\right] = \mathbf{0}$$

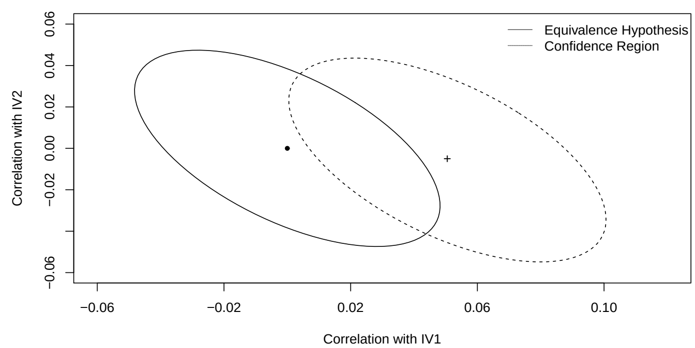
where Z_c are observations related to classroom c, $\tau^2(W_{1c}) = W'_{1c}\beta_0$ represents conditional heterogeneity in teacher effectiveness as a function of classroom-level covariates W_{1c}, γ_0 is the peers effect parameter (where $\gamma_0 = 1$ corresponds to no spillover), and W_{2c} denotes class size. I focus on results concerning math test scores as reported in Graham (2008, Table 1, Column 1). In this application, the classroom-level covariates W_{1c} are school dummy variables as well as a binary variable indicating whether classroom is of the regular with a full time teaching aide type, while W_{2c} is binary

[2]This is a slight abuse of language, since strictly speaking, $\Delta^2_{\text{inf}}(\alpha)$ determines the minimal misspecification that is not confirmed by the test.

Table 1: Equivalence tests results for social interactions model

	J	$\gamma = 2$	$\gamma = 1$	$\gamma = 0$
Test statistic	1.081	1.108	1.139	1.157
P-value ($\Delta^2 = (0.1)^2$)		0.127	0.131	0.133
$\delta^2_{\text{inf}}(5\%)$		5.557	5.649	5.70
$\Delta^2_{\text{inf}}(5\%)$		$(13.24\%)^2$	$(13.35\%)^2$	$(13.41\%)^2$

Figure 1: Social interactions: Equivalence hypothesis and confidence region in terms of correlations



indicating whether class size is small (13 to 17 students) as opposed to regular (22 to 25 students). Graham (2008) based estimation on the unconditional moments

$$\mathbb{E}\left[W_c'\,\rho(Z_c,\tau^2(W_{1c}),\gamma_0^2)\right] = \mathbf{0}\,,$$

where $W_c = (W_{1c}, W_{2c})$. To assess the approximate validity of the social interactions model, I use unconditional moments of the above type, where W_c additionally includes some interactions between binary variables. Specifically, I consider two overidentifying restrictions based on the interactions of a dummy for whether a classroom is in one of the 48 larger schools with the small and regular-with-aide class type dummies. Graham (2008) argues that such interactions terms are of particular interest if within-class-type student sorting or student-teacher matching in large schools is a potential concern.[3]

The standard two-step GMM overidentification test statistic is 1.08 and does not reject the null hypothesis that the overidentifying restrictions hold. In terms of spillovers,

[3]Considering all interactions terms of school dummies with small and regular dummies would yield a large number of restrictions with respect to the sample size $n = 317$.

the CUE-GMM estimated value of γ_0^2 is about 3.07, which is a little bit lower than the value of 3.47 reported by Graham (2008), and the p-value of a significance test of $\gamma_0^2 = 1$ (the null of no spillover) is always less than 1%. The results of the model equivalence tests for $\gamma = 2, 1$, and 0, are gathered in Table 1 and they closely agree. For $\Delta^2 = (0.1)^2$, p-values are around 13%. Thus for a significance level just above 10%, model equivalence at a tolerance $\Delta^2 = (0.1)^2$ can be accepted. The minimum tolerance that would yield to accept model equivalence for a 5% level is around $(13\%)^2$. To interpret this result, we rely on the alternative formulation of the model equivalence hypothesis

$$\mathbb{E}\, g_2'(X, \theta^*)\Sigma^{-1}\mathbb{E}\, g_2(X, \theta^*) < \Delta^2\,, \tag{3.10}$$

where, for ease of interpretation, $\mathbb{E}\, g_2(X, \theta^*)$ are the *correlations* between the error and interactions terms. Setting $\Delta^2 = (13.24\%)^2$ and estimating the matrix Σ (based on CUE-GMM results) yields an estimated set of correlations that can be confirmed by our test.[4] This set, by definition an ellipse centered at $(0, 0)$, is represented in Figure 1. The model equivalence tests at 5% level allow to conclude that the extent of misspecification is limited to correlations in this set, which include ones of 4% or less. Hence student-sorting or student-teacher matching does not appear to be of much practical importance.

It is interesting to contrast these findings with the ones that obtain from a more standard approach based on confidence regions. From estimation results, one can readily evaluate the 95% confidence region for the correlations between errors and interaction terms. This region is also represented in Figure 1. The confidence ellipse is centered at the empirical correlations. It is slightly wider than the model equivalence set and includes larger correlations values. Crucially, it does not include the point where both correlations are zero (though it would by increasing slightly the confidence level). This illustrates that confidence regions and model equivalence tests provide different information about the problem at hand.

[4]Strictly speaking, this is the largest set of correlations that is not confirmed by the test, but by a slight abuse of language, I refer to it as the smallest set that is confirmed.

3.2 Nonlinearities in Growth Regression

I consider here a cross-country growth regression in the spirit of Mankiw, Romer, and Weil (1992) using data on 86 countries averaged over the 1960's, 1970's and 1980's from King and Levine (1993) and further studied by Liu and Stengos (1999). Explanatory variables include GDP60, the 1960 level of GDP; POP, population growth (to which 0.05 is added to account for depreciation rate and technological change); SEC, the enrollment rate in secondary schools; INV, the share of output allocated to investment; and fixed time effects. The Solow model assumes a Cobb-Douglas aggregate technology, which yields a linear regression of growth on $\log(INV)$, $\log(POP)$, and $\log(SEC)$. There is more uncertainty about the relationship to the initial GDP level. Liu and Stengos (1999) argue that the relation is actually nonlinear in the initial GDP level and in human capital based on the outcome of a joint semiparametric specification test.

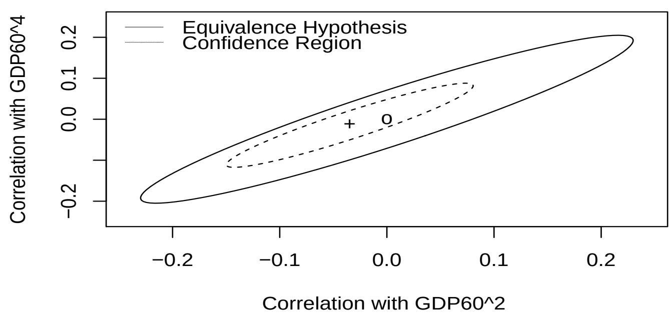
I used the proposed model equivalence tests to check whether the regression is approximately linear in the initial level of GDP and human capital. The considered restrictions are $\mathbb{E}\left(U\, W\right) = \mathbf{0}$, where U is the error term of the linear model, W contains each explanatory variable, and polynomials terms from order two to four of GDP or human capital, that is, I consider nonlinearity in initial GDP and human capital separately. In each case, there are three overidentifying restrictions. For GDP60, and when considering model equivalence at $\Delta^2 = (0.1)^2$, p-values are greater than 90%. The minimum tolerance $\Delta^2_{\text{inf}}(5\%)$ that would yield the reverse decision for a 5% level is around $(30\%)^2$. I use again the formulation in (3.10) with correlations between the error term and polynomials together with an estimated Σ to determine the smaller estimated set of correlations that can be confirmed by the model equivalence test. As this is a three-dimensional set, I report in Figure 2 a cut of this set when one of the correlation (with cubic term) is set to zero, together with the same cut of the 95% confidence region. The confidence region is much smaller than the model equivalence set and contains the point where both correlations are zero. The model equivalence set by contrast includes values larger than 20% simultaneously for both correlations.

The picture is strikingly different when considering nonlinearities in human capita. For a model equivalence test at $\Delta^2 = (0.1)^2$, all p-values are around 1%. Moreover, the minimum tolerances $\Delta^2_{\text{inf}}(5\%)$ is zero for all three tests, because all test statistics

Table 2: Equivalence tests results for growth regression

	J	$\gamma = 2$	$\gamma = 1$	$\gamma = 0$
Nonlinearities in initial GDP				
Test statistic	11.30	12.14	11.97	11.19
P-value ($\Delta^2 = (0.1)^2$)		0.93	0.92	0.90
$\delta^2_{\text{inf}}(5\%)$		23.87	23.62	22.43
$\Delta^2_{\text{inf}}(5\%)$		$(30.42\%)^2$	$(30.26\%)^2$	$(29.88\%)^2$
Nonlinearities in human capital				
Test statistic	0.203	0.222	0.223	0.224
P-value ($\Delta^2 = (0.1)^2$)		0.008	0.008	0.008
$\delta^2_{\text{inf}}(5\%)$		0	0	0
$\Delta^2_{\text{inf}}(5\%)$		0	0	0

Figure 2: Growth regression: Equivalence hypothesis and confidence region in terms of correlations



are smaller than the critical value $c_{0.05,3,0}$. This constitutes strong evidence in favor of approximate linearity of growth with respect to human capital, which is accepted at level 5% regardless of how small the tolerance is. It is noteworthy that, by contrast, a confidence region for correlations between error term and polynomials cannot be arbitrarily small, so our model equivalence hypothesis is not a confidence region of a special kind. Our finding that the model is approximately linear in $\log(SEC)$ does not actually contradict Liu and Stengos (1999). Indeed, their separable semiparametric model appears to be only slightly non-linear in $\log(SEC)$, as seen in their Figure 2, with a large confidence band that does not exclude linearity.

4 Asymptotic Properties

We rely on the concept of semiparametric power envelope and we restrict to tests that are invariant to linear transformations of the moment restrictions and of the parameters. We consider a sufficiently rich family of parametric distributions for the unknown data generating process that are differentiable in quadratic mean. For the asymptotic equivalent experiment, see Le Cam and Lo Yang (2000) and van der Vaart (1998), we determine an upper bound for the power of any invariant test using a result in Lavergne (2014), and show that our tests asymptotically attain this bound. Formally, consider the partition of $g(\cdot,\theta)$ into the p-vector $g_1(\cdot,\theta) = Lg(\cdot,\theta)$ and the remaining $(m-p)$ vector $g_2(\cdot,\theta) = Mg(\cdot,\theta)$, where $[L, M]$ is full rank. Let $\lambda = (\theta', \upsilon')' \in \Lambda = \Theta \times \mathbb{R}^{m-p}$, and define

$$h(X,\lambda) = \begin{bmatrix} g_1(X,\theta) \\ g_2(X,\theta) - \upsilon \end{bmatrix}.$$

We consider the following family of probability distributions.

Definition 4.2 *$\mathcal{P}$ is a family of probability distributions P_λ, $\lambda \in \Lambda$, with common support and such that $\mathbb{E}_{P_\lambda} h(X,\lambda) = \mathbf{0}$. It contains at least one distribution with $\bar\lambda = (\bar\theta', \mathbf{0}')'$, where $\bar\theta \in \overset{\circ}{\Theta}$, the interior of Θ. The corresponding density (or probability mass function) is differentiable with respect to λ for any x, and the density and its derivatives are dominated over Λ by an integrable function. The family $\mathcal{P}$ is differentiable in quadratic mean and the limiting information matrix is $J = H'V^{-1}H$, where $H = \nabla_{\lambda'}\mathbb{E}_{P_\lambda} h(X,\lambda)$, and $V = \mathrm{Var}_{P_\lambda} h(X,\lambda)$.*

Such a family of distributions can generally be built as multinomial distributions, see Chamberlain (1987) who uses such a construct to study asymptotic efficiency bounds. In specific models, one can consider a more adapted family of distributions, see Gourieroux and Monfort (1989, Chap. 23). It is also possible to consider a family of distributions indexed by a parameter of higher dimension without affecting the analysis.

The following result shows that the model equivalence tests attain the local asymptotic power envelope of tests of H_{0n} against H_{1n} for any parametric sub-family of models $\mathcal{P}$. Here local means that we are considering parameters value around $\bar\lambda = (\bar\theta', \mathbf{0}')'$. Formally the set $\partial H_{1n}(\nu)$ introduced below allows to focus on alternatives distant enough from the null hypothesis for which power is not trivial. The result obtains independently of the specific value of $\bar{\theta}$ or the precise form of the distributions P_λ. We consider the following supplementary assumption, that corresponds to the technical conditions in Broniatowski and Keziou (2012) for asymptotics of GEL estimators under misspecification, see Newey and Smith (2004) for the case of a well specified model.

Assumption 4.4 *(a)* $\mathbb{E}\,\sup_{\theta\in\Theta}\|g(X,\theta)\|^\alpha < \infty$ *for some* $\alpha > 2$
(b) Let $m_\gamma(X,\theta,t) = t_0 - \psi_\gamma\left(t_0 + \sum_{l=1}^m t_l g_l(X,\theta)\right)$.
Then $\tilde{\theta}_0 = \arg\inf_\Theta \sup_{\mathcal{T}_\theta} \mathbb{E}\, m_\gamma(X,\theta,t)$, *where* $\mathcal{T}_\theta = \{t \in \mathbb{R}^{1+m} : \mathbb{E}\,|m_\gamma(X,\theta,t)| < \infty\}$, *exists, is unique, and belongs to* $\overset{\circ}{\Theta}$. *Moreover, for some neighborhood* $N_{\tilde{\theta}_0}$ *of* $\tilde{\theta}_0$, $\mathbb{E}\,\sup_{\theta\in N_{\tilde{\theta}_0}}\|\nabla_\theta g(X,\theta)\| < \infty$.
(c) Let $\bar{t}(\theta) = \sup_{\mathcal{T}_\theta}\mathbb{E}\, m_\gamma(X,\theta,t)$. *Then* $\mathbb{E}\,\sup_{\theta\in\Theta}\sup_{t\in N_{\bar{t}(\theta)}}|m_\gamma(X,\theta,t)| < \infty$, *where* $N_{\bar{t}(\theta)} \subset \mathcal{T}_\theta$ *is a compact set such that* $\bar{t}(\theta) \in \overset{\circ}{N}_{\bar{t}(\theta)}$.

Theorem 4.1 *Suppose* $X_1,\ldots,X_n$ *are i.i.d. according to* $P_\lambda \in \mathcal{P}$ *as defined above, and that Assumptions 2.1, 2.2, 2.3, and 4.4 hold.*

(A) Let ϕ_n *be a pointwise asymptotically level* α *tests sequence, that is*

$$\limsup_{n\to\infty}\mathbb{E}_{P_\lambda}(\phi_n) \le \alpha \quad \forall P_\lambda \in H_{0n}\cap\mathcal{P}\,.$$

Let $M > 0$ *arbitrary large and* $\mathcal{N}(\bar{\lambda},M) = \left\{\bar{\lambda} + n^{-1/2}\Upsilon,\ \Upsilon\in\mathbb{R}^m,\ \|\Upsilon\|\le M\right\}$. *If* ϕ_n *is invariant to orthogonal transformations of the parameters and of the moment restrictions, then for all* $\nu^2 < \delta^2$

$$\limsup_{n\to\infty}\mathbb{E}_{P_\lambda}(\phi_n) \le \Pr\left[\chi^2_r(\nu^2) < c_{\alpha,r,\delta^2}\right] \quad \forall P_\lambda \in \partial H_{1n}(\nu)\cap\mathcal{P}\,, \quad \lambda\in\mathcal{N}(\bar{\lambda},M)\,, \tag{4.11}$$

where $\partial H_{1n}(\nu) = \{P_\lambda : 2\,D_\gamma(\mathcal{M},P_\lambda) = \nu^2/n\}$.
(B) The tests sequence π_n *is pointwise asymptotically level* α *for any* $P_\lambda \in H_{0n}\cap\mathcal{P}$ *with* $\lambda\in\mathcal{N}(\bar{\lambda},M)$, *is invariant to orthogonal transformations of the parameters and of the moment restrictions, and is such that for all* $\nu^2<\delta^2$

$$\limsup_{n\to\infty}\mathbb{E}_{P_\lambda}(\pi_n) = \Pr\left[\chi^2_r(\nu^2) < c_{\alpha,r,\delta^2}\right] \quad \forall P_\lambda \in \partial H_{1n}(\nu)\cap\mathcal{P}\,, \quad \lambda\in\mathcal{N}(\bar{\lambda},M)\,.$$

Our model equivalence test attains the power envelope of tests of H_{0n} that are invariant to orthogonal transformations. But tests that are also invariant to possibly nonlinear transformations cannot be more powerful. Hence our test asymptotically reaches the semiparametric power envelope of invariant tests.

5 Concluding Remarks

We have proposed a new theoretical framework to assess the *approximate* validity of overidentifying moment restrictions. Approximate validity is evaluated through a Cressie-Read divergence between the true probability measure and the closest measure that imposes the moment restrictions of interest. The considered *alternative* hypothesis states that the divergence is smaller than some user-chosen tolerance. A model equivalence test is built on the corresponding empirical divergence, and attains the local semiparametric power envelope of invariant tests. Using two empirical applications, we have illustrated the usefulness of our approach, discussed how the choice of the tolerance can be adapted to the application at hand, and show how this can provide complementary information on potential misspecification compared to standard procedures.

One may be interested in assessing the approximate validity of only a subset of the moment restrictions, such as when doubt surrounds the exogeneity of some instruments. It is likely that our approach generalizes to this setup. Another direction of research could focus on a subvector of parameters of interest. This is a different issue from the one considered here, because misspecification of the model, i.e. invalid moment restrictions, can have different consequences for each parameter, and may make no difference asymptotically for some. These empirically relevant extensions are left for future research.

6 Proofs

We use the following notations. For a real-valued function $l(x, \cdot)$, $\nabla l(x, \cdot)$ and $\nabla^2 l(x, \cdot)$ respectively denote the column vector of first partial derivatives and the matrix of second derivatives with respect to its second vector-valued argument. We use indices for derivatives with respect to specific arguments.

Preliminaries:

Let $\psi_\gamma(y) = \sup_x \{yx - \varphi_\gamma(x)\}$ be the so-called convex conjugate of $\varphi_\gamma(\cdot)$. For the

Cressie-Read family of functions, the convex conjugates are

$$\begin{aligned}
\psi_\gamma(y) &= \gamma^{-1}\left[(\gamma y - y + 1)^{\frac{\gamma}{\gamma-1}} - 1\right], \quad \gamma \in \mathbb{R}\backslash\{0,1\} \\
\psi_1(y) &= \exp(y) - 1\,, \\
\psi_0(y) &= -\log(1-y)\,,
\end{aligned}$$

where the domain may vary depending on γ. By definition, the convex conjugate is strictly convex on its domain, and due to our definition, $\psi_\gamma(0) = 0$, $\psi_\gamma^{'}(0) = 1$, and $\psi_\gamma^{''}(0) = 1$. For $t \in \mathbb{R}^{m+1}$ let $m_\gamma(X,\theta,t) = t_0 - \psi_\gamma\left(t_0 + \sum_{l=1}^m t_l g_l(X,\theta)\right)$. Duality applies provided Assumption 2.1 and 3.1 hold, see Keziou and Broniatowski (2006) and Broniatowski and Keziou (2012), and implies that

$$D_\gamma(\mathcal{M}, P) = \inf_\Theta \sup_{t\in\mathbb{R}^{m+1}} \mathbb{E}\, m_\gamma(X,\theta,t) \tag{6.12}$$

$$\text{and} \qquad D_\gamma(\mathcal{M}_n, P_n) = \inf_\Theta \sup_{t\in\mathbb{R}^{m+1}} \mathbb{E}_n\, m_\gamma(X,\theta,t)\,. \tag{6.13}$$

We now detail some key properties that will be used in our proofs. We let $\tilde{g}(X,\theta) = (\mathbb{I}(X\in\mathbb{R}^p), g'(X,\theta))'$ so that $m_\gamma(X,\theta,t) = t_0 - \psi_\gamma(t'\tilde{g}(X,\theta))$, where $t = (t_0, t_1, \ldots t_m)'$.

a. $\mathbb{E}\, m_\gamma(X,\cdot,\cdot)$ is twice continuously differentiable in $t \in \mathcal{T}_\theta$ and in θ. This comes from Assumption 2.2 and the differentiability of Cressie-Read divergences.

b. It is also strictly concave in t for all θ since $\psi(\cdot)$ is strictly convex.

c. $\mathbb{E}\, m_\gamma(X,\theta,\mathbf{0}) = \mathbf{0}$,

$$\begin{aligned}
\nabla\mathbb{E}\, m_\gamma(X,\theta,\mathbf{0}) &= \begin{bmatrix} \mathbf{0} \\ 0 \\ -\mathbb{E}\, g(X,\theta) \end{bmatrix}, \\
\nabla^2\mathbb{E}\, m_\gamma(X,\theta,\mathbf{0}) &= \begin{bmatrix} \mathbf{0} & -\mathbb{E}\,\nabla_\theta\tilde{g}(X,\theta) \\ \cdot & -\mathbb{E}\,\tilde{g}(X,\theta)\tilde{g}'(X,\theta) \end{bmatrix}.
\end{aligned}$$

From Assumption 5.1, recall that $\bar{t}(\theta) = \sup_{\mathcal{T}_\theta} \mathbb{E}\, m_\gamma(X,\theta,t)$.

a. The function $\bar{t}(\cdot)$ is well-defined. Existence for any θ is ensured by Assumptions 2.1 and 3.1. By Assumption 2.2, $\operatorname{Var} g(X,\theta)$ is positive definite, and hence the functions in $g(X,\theta)$ are linearly independent, so uniqueness is ensured, see e.g. Keziou and Broniatowski (2006).

b. The function $\bar{t}(\cdot)$ is continuous and twice differentiable on Θ by the properties of $\psi_\gamma(\cdot)$ and $g(X, \cdot)$.

c. $\bar{t}(\cdot)$ admits at most one root. Indeed, $\bar{t}(\theta) = \mathbf{0} \Rightarrow \sup_T \mathbb{E}\, m_\gamma(X, \theta, t) = 0 \Rightarrow D_\gamma(\mathcal{M}_\theta, P) = 0 \Rightarrow \mathbb{E}\, g(X, \theta) = \mathbf{0} \Rightarrow \theta = \theta^*$ for a unique θ^* by Assumption 2.1.

d. Conversely, if there exists θ^* such that $\mathbb{E}\, g(X, \theta^*) = \mathbf{0}$, then $\bar{t}(\theta^*) = \mathbf{0}$. This is because on the one hand, $\mathbb{E}\, m_\gamma(X, \theta^*, \bar{t}(\theta^*)) = \sup_T \mathbb{E}\, m_\gamma(X, \theta^*, t) = 0$, and on the other hand, $\mathbb{E}\, m_\gamma(X, \theta^*, \mathbf{0}) = 0$, $\nabla_t \mathbb{E}\, m_\gamma(X, \theta^*, \mathbf{0}) = \mathbf{0}$, and $\mathbb{E}\, m_\gamma(X, \theta^*, t)$ is strictly concave in t.

Proof of Lemmas 2.1 and 3.1: We show the two lemmas in a compact way. We note that Assumption 3.1(a) is automatically satisfied for the chi-square divergence because we consider signed measures, and that this condition ensures that duality applies, see Broniatowski and Keziou (2012), Keziou and Broniatowski (2006).

Let $\lambda = (\theta', \upsilon')' \in \Lambda = \Theta \times \mathbb{R}^{m-p}$, and define

$$h(X, \lambda) = \begin{bmatrix} g_1(X, \theta) \\ g_2(X, \theta) - \upsilon \end{bmatrix}.$$

Let $\tilde{h}(X, \lambda) = (\mathbb{I}(X \in \mathbb{R}^p), h'(X, \lambda))'$ and $m_\gamma(X, \lambda, t) = t_0 - \psi_\gamma\left(t'\tilde{h}(X, \lambda)\right)$, where $t = (t_0, t_1, \ldots t_m)'$. Under Assumptions 2.1, 2.2, and 3.1, there is a λ^*, unique by 2.1(c), such that

$$0 = \inf_{\Lambda} D_\gamma(\mathcal{M}_\lambda, P) = \inf_{\Lambda} \sup_t \mathbb{E}\, m_\gamma(X, \lambda, t) = \sup_t \mathbb{E}\, m_\gamma(X, \lambda^*, t) = \mathbb{E}\, m_\gamma(X, \lambda^*, \mathbf{0}).$$

Moreover, there exist λ_R^*, unique by 2.1(d), and unique $t_R^* = \bar{t}(\lambda_R^*)$ such that

$$D_\gamma(\mathcal{M}, P) = \inf_{\Theta \times 0} \sup_t \mathbb{E}\, m_\gamma(X, \lambda, t) = \sup_t \mathbb{E}\, m_\gamma(X, \lambda_R^*, t) = \mathbb{E}\, m_\gamma(X, \lambda_R^*, t_R^*). \quad (6.14)$$

(i). If $D_\gamma(\mathcal{M}, P) = o(1)$, then $0 = \mathbb{E}\, m_\gamma(X, \lambda_R^*, \mathbf{0}) \leq \mathbb{E}\, m_\gamma(X, \lambda_R^*, t_R^*) = o(1)$, and it follows that $\|t_R^*\| = o(1)$ since $\mathbb{E}\, m_\gamma(X, \lambda_R^*, t)$ is twice continuously differentiable and strictly concave in t. Since $\bar{t}(\lambda^*) = \mathbf{0}$ and $\bar{t}(\cdot)$ is continuous and admits only one root, it must be that $\|\lambda_R^* - \lambda^*\| = o(1)$. By a Taylor expansion of $\mathbb{E}\, m_\gamma(X, \lambda, t)$ and using the continuity of $\nabla^2 \mathbb{E}\, m_\gamma(X, \lambda, t)$ for $\|t\| = o(1)$, we obtain that uniformly in (λ, t) in a $o(1)$ neighborhood of $(\lambda^*, \mathbf{0})$

$$\mathbb{E}\, m_\gamma(X, \lambda, t) = \left[- (\lambda - \lambda^*)' \nabla_\lambda \mathbb{E}\, \tilde{h}(X, \lambda^*) t - \frac{1}{2} t' \mathbb{E}\, \tilde{h}(X, \lambda^*) \tilde{h}'(X, \lambda^*) t \right] (1 + o(1)) \,. \tag{6.15}$$

We can then solve for $\bar{t}(\lambda)$ to get

$$\sup_t \mathbb{E}\, m_\gamma(X, \lambda, t) = \frac{1}{2} (\lambda - \lambda^*)' J (\lambda - \lambda^*) (1 + o(1)) \,, \tag{6.16}$$

with $J = J(\lambda^*) = H(\lambda^*)' \operatorname{Var}^{-1} (h(X, \lambda^*)) H(\lambda^*)$ and $H(\lambda^*) = \nabla_{\lambda'} \mathbb{E}\, h(X, \lambda^*)$. Solving (6.16) for λ_R^* under the constraint $R'\lambda = [\mathbf{0}, \mathbf{I}_{m-p}]\lambda = \mathbf{0}$ yields

$$\lambda_R^* = J^{-1/2} [\mathbf{I} - P] J^{1/2} \lambda^* (1 + o(1)) \,,$$

$$D_\gamma(\mathcal{M}, P) = \frac{1}{2} \lambda^{*\prime} J^{1/2} P J^{1/2} \lambda^* (1 + o(1)) = \frac{1}{2} \upsilon \Sigma^{-1} \upsilon (1 + o(1)) = D_W(\mathcal{M}, P)(1 + o(1)) \,,$$

where

$$\Sigma = R' J^{-1} R \quad \text{and} \quad P = J^{-1/2} R [R' J^{-1} R]^{-1} R' J^{-1/2}. \tag{6.17}$$

(ii). Assume now instead that $D_W(\mathcal{M}_\lambda, P) = o(1)$, then $\|\mathbb{E}\, g(X, \theta^*)\| = o(1)$ and

$$0 \leq D_2(\mathcal{M}, P) \leq \frac{1}{2} \mathbb{E}\ (g'(X, \theta^*)) [\operatorname{Var}\, g(X, \theta^*)]^{-1}\, \mathbb{E}\ (g(X, \theta^*)) = o(1) \,.$$

So there exists $\left(\lambda_{R,2}^*, t_{R,2}^* = \bar{t}(\lambda_{R,2}^*)\right)$ such that $D_2(\mathcal{M}, P) = \mathbb{E}\, m_2\left(X, \lambda_{R,2}^*, t_{R,2}^*\right) = o(1)$. Reasoning as above, $\|\lambda_{R,2}^* - \lambda_2^*\| = o(1)$, $\|t_{R,2}^*\| = o(1)$, and $D_2(\mathcal{M}, P) = D_W(\mathcal{M}_\lambda, P)(1 + o(1))$. For any γ,

$$0 = \mathbb{E}\, m_\gamma\left(X, \lambda_{R,\gamma}^*, \mathbf{0}\right) \leq \mathbb{E}\, m_\gamma\left(X, \lambda_{R,\gamma}^*, t_{R,\gamma}^*\right) \leq \mathbb{E}\, m_\gamma\left(X, \lambda_{R,2}^*, t_{R,2}^*\right)$$

which is an $o(1)$ by a Taylor expansion at $(\lambda_2^*, \mathbf{0})$. Reason then as above to obtain $D_\gamma(\mathcal{M}, P) = D_W(\mathcal{M}_\lambda, P)(1 + o(1)) = D_2(\mathcal{M}, P)(1 + o(1))$.

Proof of Theorem 5.1:

(i). Recall that with $J = J(\lambda^*) = H(\lambda^*)' V(\lambda^*)^{-1} H(\lambda^*)$, $H(\lambda^*) = \nabla_{\lambda'} \mathbb{E}\, h(X, \lambda^*)$, and $V(\lambda^*) = \operatorname{Var} h(X, \lambda^*) = \operatorname{Var} g(X, \theta^*)$. The proof of Lemma 3.1 yields that $2\, D_\gamma(\mathcal{M}, P) = \lambda^{*\prime} J^{1/2} P J^{1/2} \lambda^* (1 + o(1))$, uniformly in $\lambda^* \in \mathcal{N}(\bar{\lambda}, M)$, where P is defined in (6.17). Moreover, and also uniformly in $\lambda^* \in \mathcal{N}(\bar{\lambda}, M)$, we have $J =$

$J(\bar{\lambda}) + o(1) = \bar{J} + o(1)$, and similarly $P = \bar{P} + o(1)$ with self-explanatory notations. Since $\bar{P}\bar{J}^{1/2}\bar{\lambda} = \bar{J}^{-1/2}R[R'\bar{J}^{-1}R]^{-1}R'\bar{\lambda} = \mathbf{0}$,

$$\begin{aligned} 2\,D_\gamma(\mathcal{M}, P) &= \lambda^{*\prime}\bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\lambda^*(1 + o(1)) = \left(\lambda^* - \bar{\lambda}\right)'\bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\left(\lambda^* - \bar{\lambda}\right)(1 + o(1)) \\ &= n^{-1}\Upsilon'\bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon(1 + o(1))\,. \end{aligned} \tag{6.18}$$

Let $\widehat{\lambda}$ be the minimum empirical divergence estimator of λ^*, that is the argument minimizing $2\inf_\Lambda\, D(\mathcal{M}_{\lambda,n}, P_n)$. Using a reasoning similar to Lemma 3.1's proof for the empirical problem yields

$$2n\,D(\mathcal{M}_n, P_n) = n\widehat{\lambda}'J_n^{1/2}P_nJ_n^{1/2}\widehat{\lambda}(1 + o_p(1)) \tag{6.19}$$

with $P_n = J_n^{-1/2}R[R'J_n^{-1}R]^{-1}R'J_n^{-1/2}$, $J_n = H_n'V_n^{-1}H_n$, $H_n = \nabla_{\lambda'}\mathbb{E}_n h(X, \widehat{\lambda})$, and $V_n = \operatorname{Var}_n g(X, \widehat{\theta})$.

(ii). If we assume correct specification of the moment restrictions, that is $\lambda = \bar{\lambda} = \left(\bar{\theta}, \mathbf{0}\right)$, standard tools, see e.g. Newey and Smith (2004, Theorem 3.2) or Broniatowski and Keziou (2012, Theorem 5.6), yield that under Assumptions 2.1, 2.2, 3.1, and 5.1,

$$\sqrt{n}\left(\widehat{\lambda} - \bar{\lambda}\right) = -\bar{J}^{-1}\bar{H}'\bar{V}^{-1}\sqrt{n}\mathbb{E}_n h(X, \bar{\lambda}) \xrightarrow{d} N(\mathbf{0}, \bar{J}^{-1})\,,$$

where $\bar{J} = J(\bar{\lambda})$, and similarly for $\bar{H}$ and $\bar{V}$. Moreover, $J_n = \bar{J} + o_p(1)$ and $P_n = \bar{P} + o(1)$. Let us now look at the behavior of $\widehat{\lambda}$ under local misspecification. Local asymptotic normality of the log-likelihood ratio, which follows as the model is differentiable in quadratic mean over Λ, see van der Vaart (1998, Theorem 7.2), yields

$$\begin{aligned} n^{1/2}\ln\prod_{t=1}^n \frac{f(X_i;\lambda)}{f(X_i;\bar{\lambda})} &= \left(\lambda - \bar{\lambda}\right)'\Delta_n - \left(\lambda - \bar{\lambda}\right)'\bar{J}\left(\lambda - \bar{\lambda}\right)/2 + o_p(1) \qquad \forall\lambda\,, \\ \text{with} \qquad \Delta_n &= n^{-1/2}\sum_{i=1}^n \nabla_\lambda \log f(X_i;\bar{\lambda}) \xrightarrow{d} N(0, \bar{J})\,, \\ \bar{J} &= \mathbb{E}\,\nabla_\lambda \log f(X;\bar{\lambda})\nabla_\lambda' \log f(X;\bar{\lambda}) = \bar{H}'\bar{V}^{-1}\bar{H}\,. \end{aligned}$$

Since $\mathbb{E}\,h\left(X, \bar{\lambda}\right) = \mathbf{0}$, total differentiation yields

$$\operatorname{Cov}\left(h\left(X, \bar{\lambda}\right), \nabla_\lambda \log f(X;\bar{\lambda})\right) = -\nabla_\lambda\mathbb{E}\,h(X, \bar{\lambda})\,.$$

Hence,
$$\begin{aligned} \operatorname{Cov}\left(\sqrt{n}\left(\widehat{\lambda} - \bar{\lambda}\right), \Delta_n\right) &= -\,n\,\bar{J}^{-1}\bar{H}'\bar{V}^{-1}\operatorname{Cov}\left(\mathbb{E}_n h(X, \bar{\lambda}), \mathbb{E}_n\nabla_\lambda \log f(X;\bar{\lambda})\right) \\ &= -\bar{J}^{-1}\bar{H}'\bar{V}^{-1}\bar{H} = -\mathbf{I}_m\,. \end{aligned} \tag{6.20}$$

Therefore by Le Cam's third Lemma, see e.g. van der Vaart (1998), we obtain that under the sequences of distributions corresponding to $\lambda = \bar{\lambda} + n^{-1/2}\Upsilon$,

$$\tau_n \equiv \sqrt{n}\left(\widehat{\lambda} - \bar{\lambda}\right) \equiv Z + o_p(1)\,,$$

where $Z \sim N(-\Upsilon, \bar{J}^{-1})$. As a consequence,

$$n\left(\widehat{\lambda} - \bar{\lambda}\right)' J_n^{1/2} P_n J_n^{1/2} \left(\widehat{\lambda} - \bar{\lambda}\right)' = Z' \bar{J}^{1/2} \bar{P} \bar{J}^{1/2} Z + o_p(1)\,.$$

(iii). Since the sequence of distributions converges to a limiting normal experiment Z with unknown mean $-\Upsilon$ and *known* covariance matrix $\bar{J}^{-1}$, it follows that we can approximate pointwise the power of any test ϕ_n by the power of a test in the limit experiment, see van der Vaart (1998, Theorem 15.1) and Lehmann and Romano (2005, Theorem 13.4.1).

Lemma 6.3 (Lavergne (2014, Lemma 4.2)) *Consider testing*

$$H_0 \;: \mu'\Omega^{-1/2}P\Omega^{-1/2}\mu \geq \delta^2 \qquad \text{against} \qquad H_1 \;: \mu'\Omega^{-1/2}P\Omega^{-1/2}\mu < \delta^2\,,$$

where P is a known orthogonal projection matrix of rank r, from one observation $Z \in \mathbb{R}^p$ distributed as a multivariate normal $N(\mu, \Omega)$ with unknown mean μ and known nonsingular covariance matrix Ω. Then the test $\pi(z)$ that rejects H_0 when $Z'\Omega^{-1/2}P\Omega^{-1/2}Z < c_{\alpha,r,\delta^2}$ is of level α. For any $\nu^2 < \delta^2$, the test is maximin among α-level tests of H_0 against $H_1(\nu)$: $\mu'\Omega^{-1/2}P\Omega^{-1/2}\mu \leq \nu^2$ with guaranteed power $\Pr\left[\chi^2_r(\nu^2) < c_{\alpha,r,\delta^2}\right]$.

In our case, the test writes $\pi(Z) = \mathbb{I}\left[Z'\bar{J}^{1/2}\bar{P}\bar{J}^{1/2}Z < c_{\alpha,r,\delta^2}\right]$. Since the test is maximin, it is necessarily admissible and unbiased. Moreover, as it is independent of ν^2, it must be most powerful against $\Upsilon = \mathbf{0}$. Finally, as it is invariant to orthogonal transformations of the parameter space, it must be UMP invariant.

(iv). For $\lambda \in \mathcal{N}(\bar{\lambda}, M)$, the model equivalence test π_n is asymptotically equivalent to $\pi(\tau_n)$, where $\pi(\cdot)$ is the test defined above and $\tau_n \equiv \sqrt{n}\left(\widehat{\lambda} - \bar{\lambda}\right)$. It thus remains to check that π_n has the same local asymptotic properties as the optimal test $\pi(Z)$ in the limiting experiment.

We have $\mathbb{E}\,\pi_n = \mathbb{E}\,\pi(\tau_n) + o(1)$ pointwise in $\Upsilon \in \mathbb{R}^m$. Also $n\tau_n' J_n^{1/2} P_n J_n^{1/2}\tau_n$ is for any Υ asymptotically distributed as a non-central $\chi^2_{m-p}(\Upsilon\bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon)$, see Rao and

Mitra (1972). As $\pi(\tau_n)$ rejects H_{0n} when $\tau_n J_n^{1/2} P_n J_n^{1/2} \tau_n < c_{\alpha,r,\delta^2}$,

$$\begin{aligned}\mathbb{E}_{\bar{\lambda}+n^{-1/2}\Upsilon}\pi(\tau_n) &= \mathbb{P}_{\bar{\lambda}+n^{-1/2}\Upsilon}\left[\tau_n' J^{1/2} P J^{1/2} \tau_n < c_{\alpha,r,\delta^2}\right] \\ &\to \mathbb{P}\left[\chi_r^2(\Upsilon \bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon) < c_{\alpha,r,\delta^2}\right] .\end{aligned}$$

Hence, $\pi(\tau_n)$ and thus π_n are locally pointwise asymptotic level α.

The proof of Lemma 6.3 in Lavergne (2014) shows that $\pi(Z)$ is a α-level Bayes test of

$$H_0 : \ \Upsilon' \bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon \geq \delta^2 \qquad \text{against} \qquad H_1(\nu) : \ \Upsilon' \bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon \leq \nu^2$$

for $\nu^2 < \delta^2$ under least favorable a priori measures, which are respectively the uniform measure Q_δ on the domain $S(\delta)$ such that $\Upsilon' \bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon = \delta^2$ and the uniform measure Q_ν defined similarly. Now

$$\mathbb{E}_{Q_\nu}\pi(\tau_n) = \int_{S(\nu)} \mathbb{E}\,\pi(\tau_n)\, dQ_\nu \to \mathbb{E}_{Q_\nu}\pi(Z)$$

by the Lebesgue dominated convergence theorem, so that $\pi(\tau_n)$ and thus π_n are also asymptotically Bayesian level α for the same a priori measures. For any other test sequence ϕ_n of asymptotically Bayesian level α,

$$\limsup_{n\to\infty} \inf_{H_1(\nu)} \mathbb{E}\,\phi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\nu}\phi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\nu}\pi(\tau_n) .$$

But $\limsup_{n\to\infty} \mathbb{E}_{Q_\nu}\pi(\tau_n) = \mathbb{E}_{Q_\nu}\pi(Z) = \inf_{H_1(\nu)} \mathbb{E}\,\pi(Z) = \lim_{n\to\infty} \inf_{H_1(\nu)} \mathbb{E}\,\pi(\tau_n)$. Gathering results,

$$\liminf_{n\to\infty} \left(\inf_{H_1(\nu)} \mathbb{E}\,\pi(\tau_n) - \inf_{H_1(\nu)} \mathbb{E}\,\phi_n \right) \geq 0\, ,$$

which shows that $\pi(\tau_n)$ and thus π_n are locally asymptotically maximin.

Consider an invariant test sequence ϕ_n of pointwise asymptotic level α. Then for any ν and any Υ such that $\Upsilon' \bar{J}^{1/2}\bar{P}\bar{J}^{1/2}\Upsilon = \nu^2$

$$\limsup_{n\to\infty} \mathbb{E}_{\bar{\lambda}+n^{-1/2}\Upsilon}\phi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\nu}\phi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\nu}\pi(\tau_n) = \lim_{n\to\infty} \mathbb{E}_{\bar{\lambda}+n^{-1/2}\Upsilon}\pi(\tau_n)\, ,$$

so that $\pi(\tau_n)$ and thus π_n have maximum asymptotic local power among invariant tests.

Since the power of $\pi(\tau_n)$ converges to a bounded function which is continuous in Υ, limits of extrema on $H_1(\nu)$ equal limits of extrema on $H_{1n}(\nu) : \ 2D_\gamma(\mathcal{M}, P) < \nu^2/n$, using (6.18). Hence the same local asymptotic properties hold for $\pi(\tau_n)$ and thus π_n as tests of H_{0n} against $H_{1n}(\nu)$.

References

ANDREWS, D. W. K. (1989): "Power in Econometric Applications," *Econometrica*, 57, 1059–1090.

ANDREWS, I., M. GENTZKOW, AND J. SHAPIRO (2018): "On the Informativeness of Descriptive Statistics for Structural Estimates," Working Paper, Harvard University.

ANDREWS, I., M. GENTZKOW, AND J. M. SHAPIRO (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments*," *The Quarterly Journal of Economics*, 132, 1553–1592.

ANTOINE, B., H. BONNAL, AND E. RENAULT (2007): "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood," *J. Econometrics*, 138, 461–487.

ARMSTRONG, T. AND M. KOLESÁR (2018): "Sensitivity Analysis using Approximate Moment Condition Models," Working Paper, Yale University.

BERGER, J. O. AND M. DELAMPADY (1987): "Testing Precise Hypotheses," *Statistical Science*, 2, 317–335.

BERKOWITZ, D., M. CANER, AND Y. FANG (2008): "Are Nearly Exogenous Instruments reliable?" *Economics Letters*, 101, 20–23.

——— (2012): "The Validity of Instruments Revisited," *Journal of Econometrics*, 166, 255–266.

BONHOMME, S. AND M. WEIDNER (2018): "Minimizing Sensitivity to Model Misspecification," Working Paper, University of Chicago, arxiv 1807.02161.

BRONIATOWSKI, M. AND A. KEZIOU (2012): "Divergences and Duality for Estimation and Test Under Moment Condition Models," *Journal of Statistical Planning and Inference*, 142, 2554–2573.

BUGNI, F. A., I. A. CANAY, AND P. GUGGENBERGER (2012): "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80, 1741–1768.

CANER, M. (2014): "Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics," *Journal of Econometrics*, 182, 247–268.

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

CHAUSSÉ, P. (2010): "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R," *Journal of Statistical Software*, 34, 1–35.

CHETTY, R. (2012): "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply," *Econometrica*, 80, 969–1018.

Christensen, T. and B. Connault (2019): "Counterfactual Sensitivity and Robustness," Working Paper, New York University, arxiv 1904.00989v2.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *Review of Economics & Statistics*, 94, 260–272.

Cox, D. R. (1958): "Some Problems Connected with Statistical Inference," *Ann. Math. Statist.*, 29, 357–372.

Cressie, N. and T. Read (1984): "Multinomial Goodness-of-Fit Tests," *J. Roy. Statist. Soc. Ser. B*, 46, 440–464.

Dette, H. and A. Munk (1998): "Validation of Linear Regression Models," *Ann. Statist.*, 26, 778–800.

Good, I. (1981): "Some Logic and History of Hypothesis Testing," in *Philosophy in Economics*, ed. by J. Pitt, Springer, 149–174.

Gourieroux, C. and A. Monfort (1989): *Statistics and Econometric Models*, Cambridge University Press.

Graham, B. S. (2008): "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica*, 76, 643–660.

Gregory, A. W. and M. R. Veall (1985): "Formulating Wald Tests of Nonlinear Restrictions," *Econometrica*, 53, 1465–1468.

Guggenberger, P. (2012): "On The Asymptotic Size Distortion of Tests When Instruments Locally Violate The Exogeneity Assumption," *Econometric Theory*, 28, 387–421.

Hall, A. R. and A. Inoue (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114, 361 – 394.

Hansen, L.-P., J. Heaton, and A. Yaron (1996): "Finite Sample Properties of Some Alternative GMM Estimators," *J. Bus. Econom. Statist.*, 14, 262–280.

Hodges, J. and E. Lehmann (1954): "Testing the Approximate Validity of Statistical Hypotheses," *J. Roy. Statist. Soc. Ser. B*, 16, 261–268.

Hoenig, J. and D. Heisey (2001): "The Abuse of Power: the Pervasive Fallacy of Power Calculations for Data Analysis," *Amer. Statist.*, 55, 19 – 24.

Imbens, G. (1993): "One Step Estimators for Over-Identified Generalized Method of Moments Models," *Rev. Econom. Stud.*, 64, 359–383.

Imbens, G., R. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Keziou, A. and M. Broniatowski (2006): "Minimization of Divergences on Sets of Signed Measures," *Studia Sci. Math. Hungar.*, 43, 403–442.

King, R. and R. Levine (1993): "Finance and Growth: Schumpeter Might Be Right," *Quart. J. Econ.*, 108, 717–737.

Kitamura, Y. (2007): "Empirical Likelihood Methods in Econometrics: Theory and Practice," in *Advances in Economics and Econometrics: Theory and Applications, 9th World Congress, vol. 3*, ed. by R. Blundell, W. Newey, and T. Persson, Cambridge University Press.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Econometrica*, 81, 1185–1201.

Kitamura, Y. and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

Kraay, A. (2012): "Instrumental Variables Regressions With Uncertain Exclusion Restrictions: a Bayesian Approach." *Journal of Applied Econometrics*, 27, 108–128.

Lavergne, P. (2014): "Model Equivalence Tests in a Parametric Framework," *J. Econometrics*, 178, 414–425.

Le Cam, L. and G. Lo Yang (2000): *Asymptotics in Statistics*, Springer.

Leamer, E. (1998): "Things That Bother Me," *Econ. Rec.*, 64, 331–335.

Lehmann, E. and J. Romano (2005): *Testing Statistical Hypotheses*, Springer Texts in Statistics, Springer.

Liu, Z. and T. Stengos (1999): "Non-Linearities in Cross-Country Growth Regressions: a Semiparametric Approach," *J. Appl. Econometrics*, 14, 527–538.

Mankiw, N. G., D. Romer, and D. N. Weil (1992): "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 107, 407–438.

McCloskey, D. (1985): " The Loss Function Has Been Mislaid: the Rhetoric of Significance Tests," *Amer. Econ. Rev.*, 75, 201–205.

Nevo, A. and A. M. Rosen (2012): "Identification With Imperfect Instruments," *Review of Economics and Statistics*, 94, 659–671.

Newey, W. and R. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Qin, J. and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *Ann. Statist.*, 22, 300–325.

Rao, C. R. and S. K. Mitra (1972): "Generalized Inverse of a Matrix and its Applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, Berkeley, Calif.: University of California Press, 601–620.

Romano, J. (2005): "Optimal Testing of Equivalence Hypotheses," *Ann. Statist.*, 33, 1036–1047.

Rosenblatt, J. (1962): "Testing Approximate Hypotheses in the Composite Case," *Annals of Math. Statist.*, 33, 1356–1364.

Schennach, S. M. (2007): "Point Estimation with Exponentially Tilted Empirical Likelihood," *Ann. Statist.*, 35, 634–672.

Smith, R. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation," *Econ. J.*, 107, 503–519.

Stock, J. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock, New York: Cambridge University Press, 80–108.

Sueishi, N. (2013): "Identification Problem of the Exponential Tilting Estimator Under Misspecification," *Economics Letters*, 118, 509–511.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wellek, S. (2003): *Testing Statistical Hypotheses of Equivalence*, Chapman and Hall/CRC.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.